An Introduction to Quantile Regression and Applications to Expenditure Data Analysis

Lan Wang

School of Statistics, University of Minnesota

May 26, 2015

Talk outline

- Expenditure data analysis
- Introduction to quantile regression
- Two applications
  - Weighted quantile regression for health expenditure data with missing covariates
  - Predicting high spending customers with semiparametric quantile regression
- Penalized quantile regression with high-dimensional covariates

Health expenditure data

- A large proportion of health care costs is concentrated in a small portion of patients.
- We note that the VHA budget for 2012 is \$54.3 billion dollars. With the 10% of high-cost patients accounting for 70% (\$36 billion) to 80% (\$40 billion) of the total costs.
- Analysis of health care cost data is often complicated by a high level of **skewness**, **heteroscedastic variances** and the presence of **missing data**.

Health expenditure data (cont'd)

Limitations of the mean regression approach:

- Transformation of the response variable is often required when constructing the mean regression model and retransformation is needed in order to obtain direct inference on the mean cost.
- The conditional mean model focuses primarily on the marginal effects of the risk factors on the central tendency of the conditional distribution. Focusing on the marginal effects at the central tendency may substantially distort the information of interest at the tails.

Poll question

How much experience do you have with Quantile Regression?

- I'm an expert!
- I have some experience.
- I have heard of it but have not used it.
- This is the first time I've heard quantile regression.

Motivations for quantile regression

- **Regression**: to obtain a summary of the relationship between a response variable y and a set of covariates $\mathbf{x}$.
- Least squares regression captures how the **mean** of y changes with $\mathbf{x}$.
- Conditional quantile functions provide **a more complete picture** of the relationship between y and $\mathbf{x}$.
- Conditional quantiles are often of **direct interest**.

Example: birth weight data

- **Response variable**: baby's birth weight
- **Covariates**: baby's gender mother's age, race, weight gain, smoking status, education level, ...
- **Lower quantiles** of birth weight are of direct interest.
- Abreveya (2001) and Koenker and Hallock (2001): covariate effects on lower quantiles may **differ from** those on the mean or median birth weight.

Quantile regression

Let $F_Y(y|\mathbf{X}=\mathbf{x}) = P(Y \le y|\mathbf{X}=\mathbf{x})$ denote the conditional CDF of Y given $\mathbf{X}=\mathbf{x}$. The **τth conditional quantile** of Y is defined as

$$Q_Y(\tau|\mathbf{X}=\mathbf{x}) = \inf\{y : F_Y(y|\mathbf{x}) \ge \tau\}, \quad 0 < \tau < 1. \tag{1}$$

- **Linear quantile regression**

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}(\tau),$$

where $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \cdots, \beta_p(\tau))^T$ is the quantile coefficient that may depend on τ.

- Let $\epsilon = Y - Q_Y(\tau|\mathbf{x})$, then we may also write

$$Y = \mathbf{x}^T\boldsymbol{\beta}(\tau) + \epsilon, \tag{2}$$

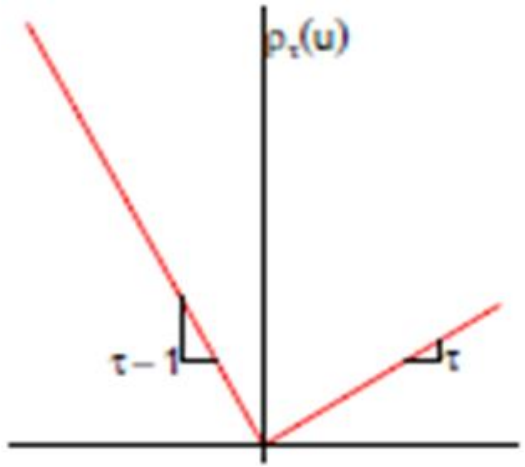
where ϵ satisfies $P(\epsilon \le 0|\mathbf{x}) = \tau$.

Calculation: an optimization perspective

- $E(Y) = \arg\min_a E\{(Y-a)^2\}$.
- Median $Q_Y(0.5) = \arg\min_a E|Y-a|$
- **Conditional quantile as a minimizer** (Koenker and Bassett, 1978):

$$\beta(\tau) = argmin_{\beta} E[\rho_\tau(Y - \mathbf{x}^T\beta)|\mathbf{X} = \mathbf{x}],$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is called the check function.



- **Computation**: linear programming (default: the simplex algorithm)

Properties of the estimator

Asymptotic normality:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)) \to N(0, \Sigma)$$

where $\Sigma = \tau(1-\tau)\big(E[f_\epsilon(0|\mathbf{x})\mathbf{x}\mathbf{x}']\big)^{-1} E(\mathbf{x}\mathbf{x}')\big(E[f_\epsilon(0|\mathbf{x})\mathbf{x}\mathbf{x}']\big)^{-1}$.

Statistical inference: kernel estimation of standard error (se="ker"), resampling-based estimation of standard error (se="boot")

R example: birth weight data

```
> library(quantreg)
> dat1<-read.table("birth.txt", header=TRUE, sep=";")
> attach(dat1)
> AGE2<-AGE^2
> summary(rq(WEIGHT~BOY+BIRTHRECORD+ BLACK+SMOKER+COLLEGE+WEIGHTGAIN
             +AGE+AGE2,tau=0.5))

tau: [1] 0.5

Coefficients:
             Value      Std. Error t value    Pr(>|t|)
(Intercept) 2912.28554  166.83340   17.45625    0.00000
BOYM          89.91828   17.70474    5.07877    0.00000
BIRTHRECORD   18.31326    5.77476    3.17126    0.00153
BLACKTRUE   -287.45878   33.03094   -8.70271    0.00000
SMOKERTRUE  -161.13871   25.84372   -6.23512    0.00000
COLLEGETRUE   16.01326   20.64735    0.77556    0.43805
WEIGHTGAIN     2.85663    0.44219    6.46023    0.00000
AGE           25.82043   12.02136    2.14788    0.03177
AGE2          -0.41816    0.20712   -2.01888    0.04355
```

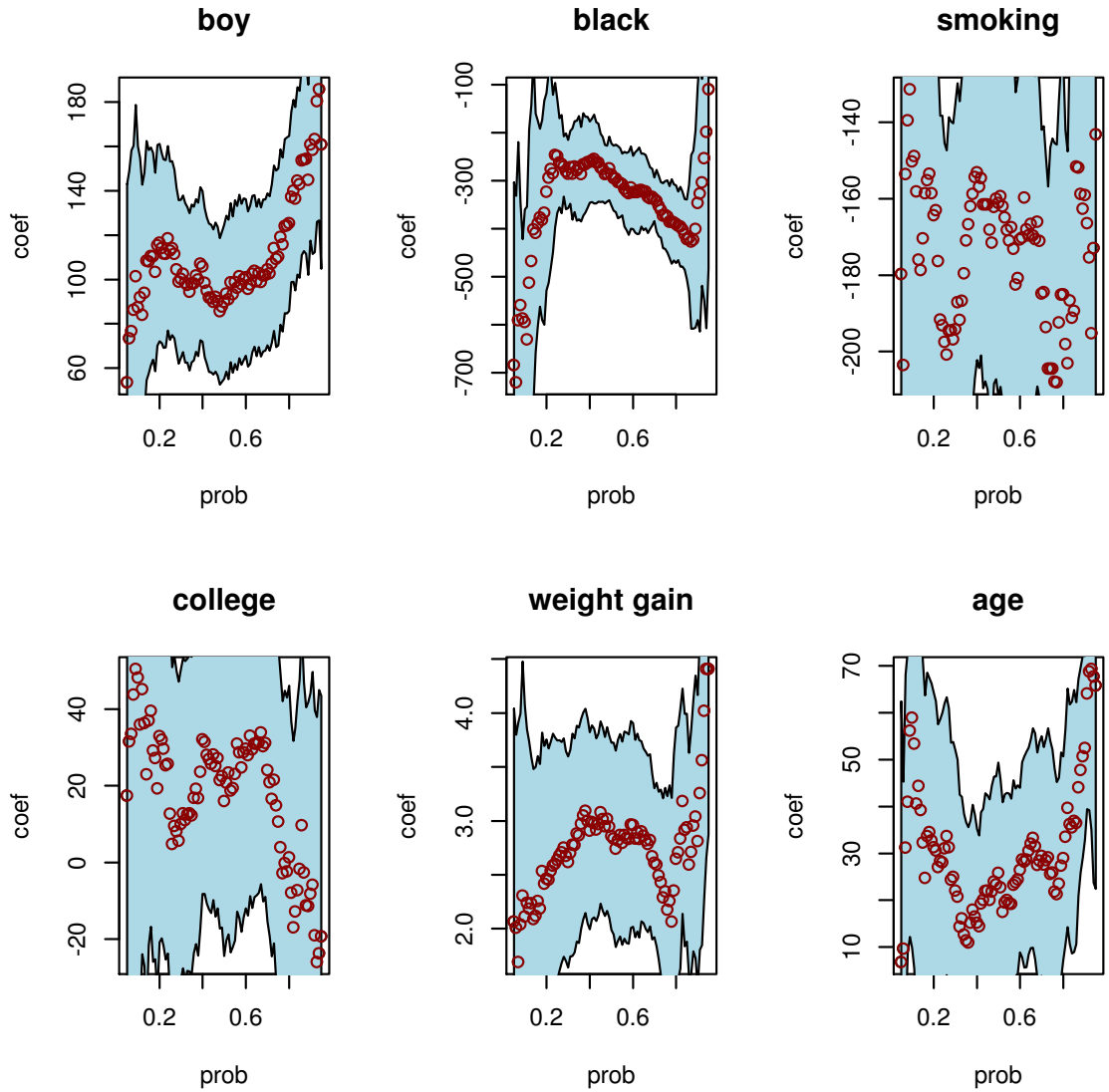
R example: birth weight data (cont'd)

```
> summary(rq(WEIGHT~BOY+BIRTHRECORD+ BLACK+SMOKER+COLLEGE+WEIGHTGAIN
             +AGE+AGE2,,tau=0.1))

Call: rq(formula = WEIGHT ~ BOY + BIRTHRECORD + BLACK + SMOKER +
      COLLEGE +WEIGHTGAIN + AGE + AGE2, tau = 0.1)
tau: [1] 0.1

Coefficients:
             Value      Std. Error t value    Pr(>|t|)
(Intercept) 1914.54960  346.03223    5.53286    0.00000
BOYM          87.33791   37.10524    2.35379    0.01862
BIRTHRECORD    2.88285   11.20978    0.25717    0.79706
BLACKTRUE   -594.96180  119.84982   -4.96423    0.00000
SMOKERTRUE  -150.24584   48.38939   -3.10493    0.00191
COLLEGETRUE   48.23029   44.10984    1.09341    0.27427
WEIGHTGAIN     2.11297    0.99950    2.11403    0.03456
AGE           58.91430   25.00547    2.35606    0.01851
AGE2          -1.11540    0.43271   -2.57772    0.00997
```

R example: birth weight data (cont'd)

```
>u=seq(.05,.95,by=.01)
>coefstd=function(u) summary(rq(WEIGHT~BOY+BIRTHRECORD+ BLACK+SMOKER
            +COLLEGE+WEIGHTGAIN+AGE+AGE2,,tau=u))$coefficients[,2]
>coefest=function(u) summary(rq(WEIGHT~BOY+BIRTHRECORD+ BLACK+SMOKER
            +COLLEGE+WEIGHTGAIN+AGE+AGE2,,tau=u))$coefficients[,1]
>CS=Vectorize(coefstd)(u)
>CE=Vectorize(coefest)(u)
>k=2
>plot(u,CE[k,],xlab="prob",ylab="coef", main="boy")
>polygon(c(u,rev(u)),c(CE[k,]+1.96*CS[k,],rev(CE[k,]-1.96*CS[k,])),
         col = "light blue")
>points(u,CE[k,], col = "dark red")
```

R example: birth weight data (cont'd)



SAS example: birth weight data (cont'd)

```
proc quantreg data=birth alpha=0.01 ci=resampling;
      model WEIGHT=BOY BIRTHRECORD BLACK SMOKER COLLEGE
                  WEIGHTGAIN AGE AGE2 / quantile=0.9
                                        CovB CorrB
                                        seed=12345;
      test_age_quadratic: AGE2 / wald lr;
```

Application I: weighted quantile regression for estimating health costs data with missing covariates (Sherwood, Wang and Zhou, 2014)

- The data (695 patients) came from a clinical study on the cost-effectiveness of a computer-assisted prospective drug utilization review program conducted in the primary care system of Indiana University Medical GroupPrimary Care.
- Response variable ("charge") is the log-transformed amount ($) charged for the health care on each of the patients.
- Seven covariates: aa (whether the patient is African-American), female, pharm_sat (pharmacist satisfaction score), alone (whether the patient is living alone), SF36_PF (SF-36 physical function score), badReaction (whether the patient stops medication because of adverse effects) and sexuallyActive (whether the patient engages in sexual activity).

Weighted quantile regression for estimating health costs data with missing covariates (cont'd)

- About 10% patients have missing values on the covariates vector (pharm_sat, SF36_PF), while all the other covariates are fully observed for all the patients.
- When the data are obtained from hospital records, incomplete records may lead to missing information. Missing data may also arise because the patients drop out of the study or are lost to follow up.
- The imputation approach often requires the specification of a joint or conditional likelihood. However, correct specification of the likelihood function is often challenging in practice, especially for skewed and heteroscedastic data or when the missing data contain both continuous and discrete variables.

Weighted quantile regression for estimating health costs data with missing covariates (cont'd)

- For subject i, $i = 1, \ldots, n$, we observe a response variable Y_i, $W_i = (W_{i1}, \ldots, W_{ip})'$ is always fully observed, and $V_i = (V_{i1}, \ldots, V_{iq})'$ may contain some missing components. Let $X_i = (W_i', V_i')'$, $R_i = 1$ if V_i is fully observed and 0 otherwise.
- Missing at random (MAR):

$$P(R_i = 1 \mid Y_i, X_i) = P(R_i = 1 \mid Y_i, W_i),$$

For an unknown γ and $T_i = (Y_i, W_i')'$, $P(R_i = 1 \mid Y_i, X_i) = \pi(T_i, \gamma)$.

- Weighted estimating equation

$$G_n^W(\beta) = \sum_{i=1}^{n} \frac{R_i}{\pi(T_i, \gamma)} \psi_\tau(Y_i - X_i'\boldsymbol{\beta}) = 0,$$

where $\Psi_\tau(t) = \tau - I(t < 0)$ is the gradient function of $\rho_\tau(t)$. To see that the weighted estimating equation is unbiased, we observe

$$\begin{aligned}
& E\left[\frac{R_i}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i'\beta(\tau))\right] \\
= \quad & E\Big[E\big[\frac{R_i}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i'\beta(\tau))\big|X_i, Y_i\big]\Big] \\
= \quad & E\left[\frac{\pi(T_i, \gamma)}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i'\beta(\tau))\right] \\
= \quad & E\Big[X_i E[\Psi_\tau(Y_i - X_i'\beta(\tau))|X_i]\Big] = 0.
\end{aligned}$$

- The estimator can be computed by weighted quantile regression

$$\widehat{\beta}_n^W = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{R_i}{\pi(T_i, \widehat{\gamma})} \rho_\tau(Y_i - X_i'\beta).$$

  The estimator is **asymptotically normal**.

- **Variable selection**: The modified BIC for the candidate model ν is defined as

$$\text{BIC}(\nu) = \min_{\beta_\nu} \left\{ \sum_{i=1}^{n} \frac{R_i}{\pi(T_i, \widehat{\gamma})} \rho_\tau(Y_i - X_{i\nu}'\beta_\nu) + \frac{d_\nu \log n}{2} \right\}. \qquad (3)$$

  where $\widehat{\gamma}$ is the estimator from the logistic regression model using all candidate covariates. The BIC procedure is **consistent**: $P(\widehat{\nu} = \nu_0) \to 1$.

Table: Analysis of health care costs data: estimation for the full model (with p-values in the parentheses)

	Weighted	Naive	Weighted0.8	Naive0.8	Weighted0.9	Naive0.9
Intercept	8.84 (0)	8.77 (0)	9.4 (0)	9.56 (0)	10.77 (0)	11.1 (0)
aa	-0.21 (0.03)	-0.19 (0.06)	-0.37 (0.03)	-0.34 (0.06)	-0.48 (0.05)	-0.38 (0.1)
female	-0.22 (0.1)	-0.26 (0.06)	-0.33 (0.11)	-0.45 (0.03)	-0.81 (0.02)	-1.05 (0)
pharm_sat	-0.2 (0.01)	-0.18 (0.03)	0 (0.99)	0.02 (0.84)	-0.06 (0.65)	-0.11 (0.42)
alone	0.1 (0.45)	0.13 (0.32)	0.52 (0.01)	0.47 (0.01)	0.47 (0.08)	0.37 (0.16)
SF36_PF	-0.01 (0)	-0.01 (0)	-0.01 (0)	-0.01 (0)	-0.01 (0)	-0.01 (0)
badReaction	0.39 (0.04)	0.36 (0.06)	0.69 (0.01)	0.64 (0.02)	0.6 (0.05)	0.66 (0.04)
sexuallyActive	-0.21 (0.05)	-0.17 (0.12)	-0.2 (0.32)	-0.21 (0.29)	-0.23 (0.36)	-0.14 (0.57)

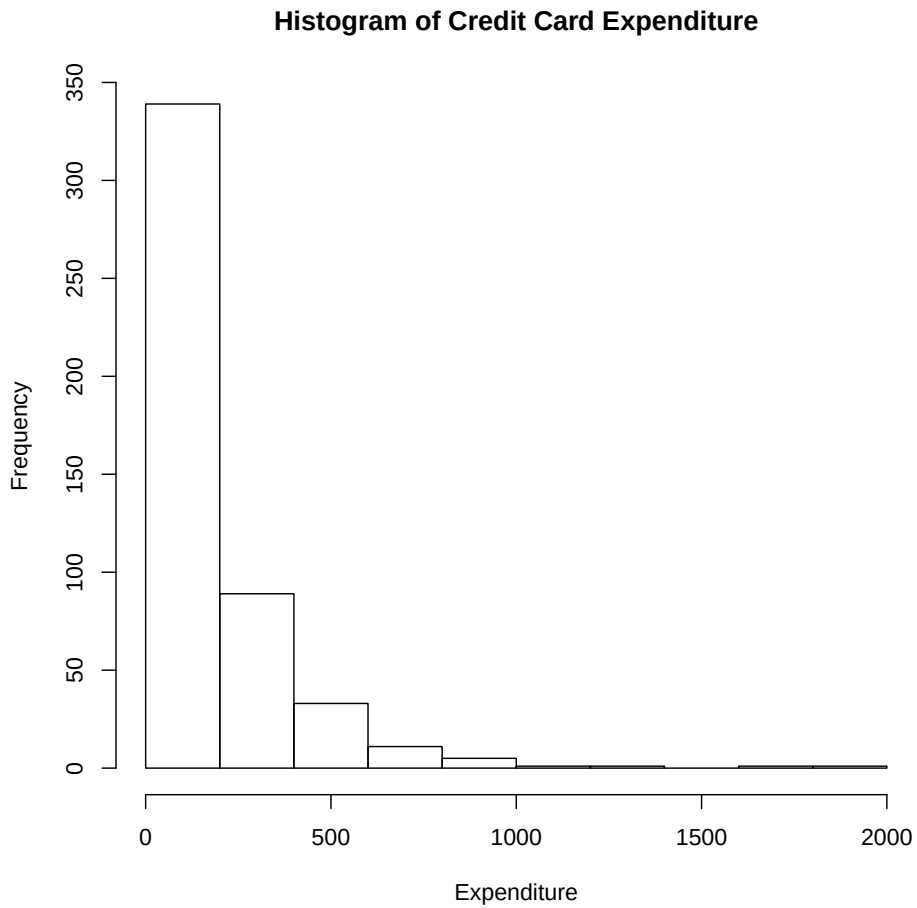
Table: Analysis of health care costs data: results from variable selection using the modified BIC at $\tau = 0.5$, 0.8 and 0.9

τ	0.5	0.80	0.90
Intercept	8.09	8.99	10.68
aa	-	-	-
female	-	-	-0.91
pharm_sat	-	-	-
alone	-	-	-
SF36_PF	-0.01	-0.01	-0.01
badReaction	-	0.77	-
sexuallyActive	-	-	-

Application II: Predicting high spending customers using semiparametric quantile regression (with Maidman, in progress)

- In many applications, it is important to predict if a future response occurs at the tails (upper tail or lower tail) of the response distribution.
- Credit card expenditure data from (Greene, 2008): identify applicants who are more likely than not to spend more than 90% of the population.
- We restrict our analysis to applicants between the ages of 18 and 28. A threshold of \$450, corresponding to the approximate 0.9 quantile of observed expenditure, was used to classify applicants as high- or low-spending. Response: $\tilde{y}_i \equiv \log(y_i + 1)$.
- The linear predictors are share, selfemp, dependents, months, owner, majorcards, and card. The nonlinear predictors are age (age in years plus twelfths of a year of the applicant) and income (yearly income in USD 10,000 of the applicant).

Predicting high spending customers using semiparametric quantile regression (cont'd)



- **Logistic regression approach**: Given a threshold (such as $5,000), which is tunable according to the goal of the study, the logistic regression approach first artificially discretizes credit card expenditure as 0-1 variables. The logistic regression model is applied to the 0-1 response data.
- The logistic regression approach has three main drawbacks:
  - Due to the artificial discretizing, there is a potential loss of information compared with the quantile regression approach, as all credit card expenditures above the threshold are treated as the same.
  - It is not clear whether the artificial 0-1 data, which are obtained by thresholding expenditure from a long-tailed and heteroscedastic distribution, would satisfy the modeling assumptions of logistic regression.
  - When prediction of expenditure is of primary concern, the logistic regression model can only predict if the expenditure exceeds a given threshold or not; but provides little information on the likely magnitude of credit card expenditure.

Partially linear additive quantile regression

- $$Q_{Y|X,Z}(\tau) = x'\beta(\tau) + \sum_{k=1}^{q} g_k(z_k),$$

  where $g_k(\cdot)$ is an unknown smooth function, $k = 1, \ldots, q$.
- Let $\pi(t) = (b_1(t), ..., b_{k_n+l+1}(t))'$ denote a vector of normalized B-spline basis functions of order $l+1$ with k_n quasi-uniform internal knots on $[0, 1]$. Then $g_k(\cdot)$ can be approximated by $\pi(z_k)'\xi_k$, $k = 1, \ldots, q$.
- $$\underset{\{\beta,\xi_1,\ldots,\xi_q\}}{\text{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \left[X_i'\beta + \sum_{k=1}^{q} \pi(z_k)'\xi_k \right] \right).$$

The new prediction method

- **Motivation**: Given a threshold c, for a new applicant with predictor vector x^*, logistic regression classifies the applicant as high-spending if $\log\left(\frac{P(Y^*>c|x^*)}{1-P(Y^*>c|x^*)}\right) > 0$ and low-spending otherwise. Or equivalently, the applicant is classified as high-spending if the estimated probability $P(Y^* > c|x^*)$ is above 0.5.
- The proposed new approach also classifies the applicant into either high-spending or not high-spending according to whether the estimated probability $P(Y^* > c|x^*)$ is above or below 0.5. However, different from the logistic regression approach, we do not need to impose a parametric distribution assumption or rely on the likelihood method. Our approach is based on the **important observation** that $P(Y^* > c|x^*) > 0.5$ is equivalent to the conditional median $Q_{Y^*|x^*}(0.5) > c$.

The new prediction method (cont'd)

- From this observation and observed data, we can classify a new applicant with predictors x^* and z^* as high- or not high-spending using a threshold c as follows:
  (a) Compute $\widehat{Q}_{y^*|x^*,z^*}(.5) = x^*\hat{\beta} + \sum_{k=1}^{q} \hat{g}_k(z_k^*)$.
  (b) If $\widehat{Q}_{y^*|x^*,z^*}(.5) > c$, classify the new applicant as high-spending; otherwise, classify as low-spending.

- $$P(Y^* \in \text{high spending} \mid Y^* > c, x^*, z^*) \to 1, \quad (4)$$
  $$P(Y^* \in \text{not high spending} \mid Y^* \leq c, x^*, z^*) \to 1, \quad (5)$$

  as $n \to \infty$.

Credit card data analysis

The predictive model used for classification is

$$\begin{aligned}\widehat{Q}_{\tilde{y}^*|x^*,z^*}(.5) &= -0.018 + 9.094 \cdot \text{share} - 0.090 \cdot \text{selfemp} - 0.053 \cdot \text{depende} \\ &\quad + 4.029 \cdot \text{card} - 0.000 \cdot \text{months} + 0.039 \cdot \text{owner} \\ &\quad + 0.020 \cdot \text{majorcards} + \hat{g}_{\text{age}}(\text{age}) + \hat{g}_{\text{income}}(\text{income}),\end{aligned}$$

Threshold	Model	FP	FN	TP	TN	ER	WE2	WE5
	Logistic	0.02	0.18			0.03	0.05	0.10
$c = 450$	PLALOG	0.02	0.13	21.53	218.47	0.03	0.04	0.08
$\tau = .9$	Quantile	0.01	0.12			0.02	0.03	0.07
	PLAQR	0.01	0.10			0.02	0.03	0.06

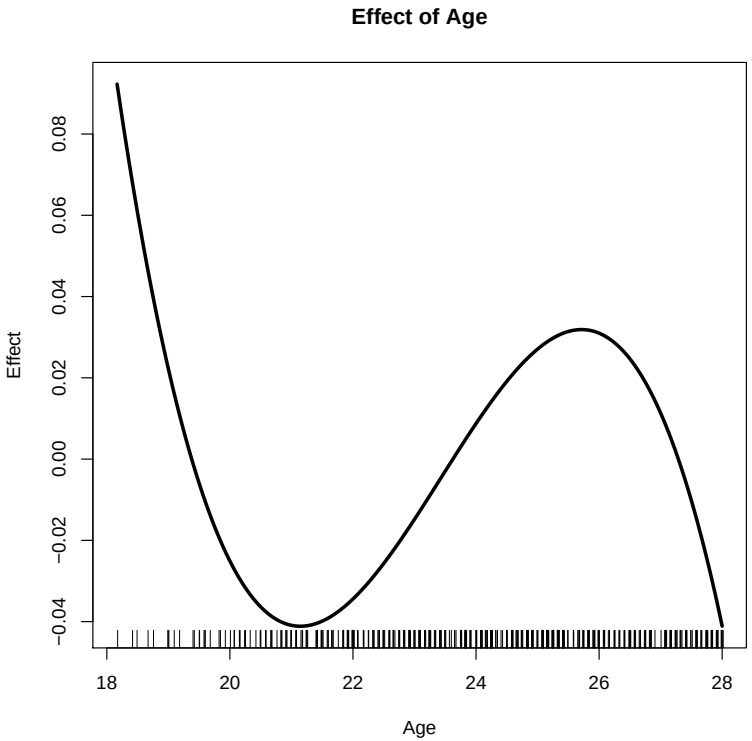


Figure: $\hat{g}_{\text{age}}$: effect plot for age.

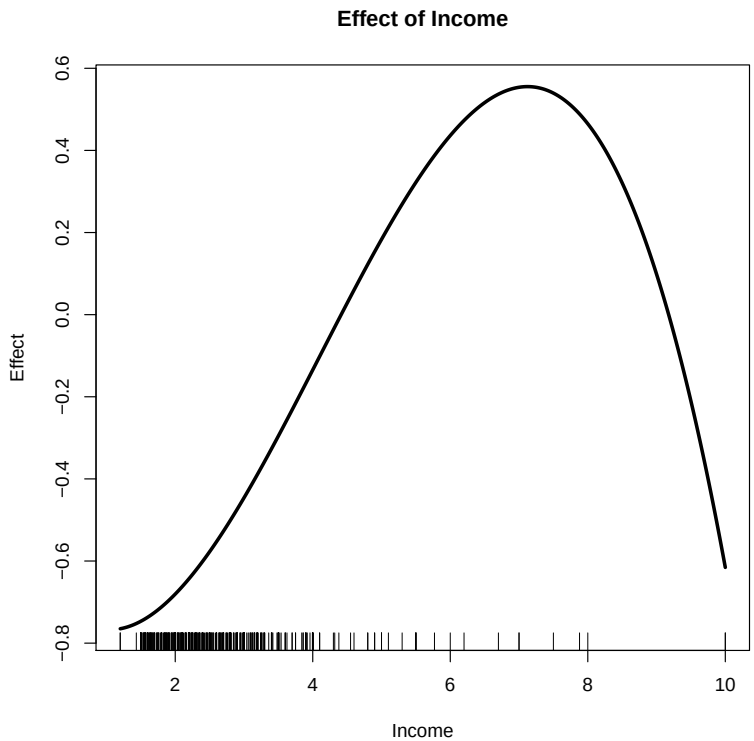


Figure: $\hat{g}_{\text{income}}$: effect plot for income.

High-dimensional data analysis

- High-dimensional data have become common in diverse fields
- **Regularization methods**
  - LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), Dantzig selector (Candes and Tao, 2007)
  - SCAD (Fan and Li, 2001; Zou and Li, 2008; Fan and Lv, 2011)
  - MCP (Zhang, 2010)
- Current literature mainly focus on **mean regression function**.

Quantile approach in high dimension

High dimension: $p \gg n$.

- **Quantile-adaptive sparsity**:
  A small number of covariates influence the conditional distribution of the response variable given all candidate covariates; however, the sets of relevant covariates may be different when we consider different segments of the conditional distribution.
- **Weaker conditions for theory**:
  No need to impose restrictive distributional or moment conditions on the random errors and allow their distributions to depend on the covariates.

Penalized linear quantile regression

- **Quantile regression (QR)**: $\widehat{\boldsymbol{\beta}}_\tau = \text{argmin}_{\boldsymbol{\beta}}\, n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta})$.
- **Penalized linear quantile regression (PQR)**: $Q(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{p}_\lambda(|\beta_\mathbf{j}|)$, where $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ.
- PQR for linear models with L_1 penalty was studied by
  - Li and Zhu (2008), Zou and Yuan (2008) for fixed p
  - Belloni and Chernozhukov (2011) for high-dimensional p_n

  PQR for linear models with nonconvex penalties was studied by
  - Wu and Liu (2009), Kai, Li and Zou (2011) for fixed p

Non-convex penalized high-dimensional PQR

Sparsity: Let $A_0 = \{j : \beta_j^* \neq 0\}$ and $|A_0| = q$. Assume that $q \ll n$.

$$Q(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

The penalty function $p_\lambda(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $\dot{p}_\lambda(t)$ on $(0, +\infty)$.

- **SCAD penalty**:

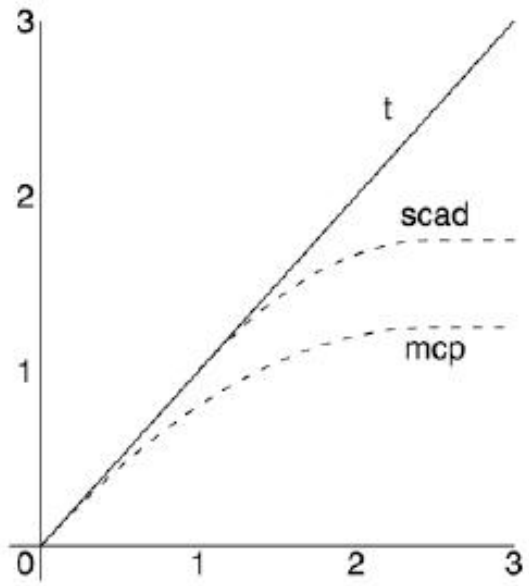
$$\begin{aligned} p_\lambda(|\beta|) &= \lambda|\beta| I(0 \leq |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \leq |\beta| \leq a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some } a > 2. \end{aligned}$$

- **MCP penalty**:

$$p_\lambda(|\beta|) = \lambda\left(|\beta| - \frac{\beta^2}{2a\lambda}\right) I(0 \leq |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \geq a\lambda), \quad a > 1.$$

Concave penalty function

- Fan and Li (2001) demonstrated that the SCAD penalty simultaneously achieves three desirable properties of penalized variable selection: **unbiasedness, sparsity and continuity**. The same properties are shared by the MCP penalty.



Difference convex program

- Difference Convex (DC) program: we consider penalized loss functions belonging to the class

$$\mathbf{F} = \{f(\mathbf{x}) : f(\mathbf{x}) = g(\mathbf{x}) - h(\mathbf{x}), \quad g, h \text{ are both convex}\}$$

  $\Rightarrow$ provides us a new formulation of oracle property.
- Extension of the KKT condition
- Difference Convex (DC) functions$\Rightarrow$ Oracle property under relaxed conditions.

Oracle property

Theorem

Assume that conditions (C1)-(C5) hold. Let $\mathcal{B}_n(\lambda)$ be the set of local minima of the nonconvex penalized quantile objective function with either the SCAD penalty or the MCP penalty and tuning parameter λ. The oracle estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ satisfies that

$$P(\widehat{\boldsymbol{\beta}} \in \mathcal{B}_n(\lambda)) \to 1$$

as $n \to \infty$ if $\lambda = o(n^{-(1-c_2)/2})$, $n^{-1/2}q = o(\lambda)$ and $\log(p) = o(n\lambda^2)$.

It can be shown that if we take $\lambda = n^{-1/2+\delta}$ for some $c_1 < \delta < \frac{c_2}{2}$, then these conditions are satisfied. We can also have $p = o(\exp(n^{\delta}))$.

Algorithm

We combine the local linear approximation algorithm (LLA, Zou and Li, 2008) with linear programming.

- While minimizing $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$, we initialize by setting $\tilde{\beta}_j^{(0)} = 0$ for $j = 1, 2, \cdots, p$. For each step $t \geq 1$, we update by solving

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \sum_{j=1}^{p} w_j^{(t-1)}|\beta_j| \right\},$$

where $w_j^{(t-1)} = p'_\lambda(|\tilde{\beta}_j^{(t-1)}|) \geq 0$ denotes the weight and $p'_\lambda(\cdot)$ denotes the derivative of $p_\lambda(\cdot)$.

Algorithm (cont'd)

- With the aid of slack variables ξ_i^+, ξ_i^-, and ζ_j, the convex optimization problem can be equivalently rewritten as

$$\min_{\boldsymbol{\xi},\boldsymbol{\zeta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tau \xi_i^+ + (1-\tau)\xi_i^-) + \sum_{j=1}^{p} w_j^{(t-1)} \zeta_j \right\}$$

$$\text{subject to} \quad \xi_i^+ - \xi_i^- = Y_i - \mathbf{x}_i^T \boldsymbol{\beta};\ i = 1, 2, \cdots, n,$$

$$\xi_i^+ \geq 0, \xi_i^- \geq 0;\ i = 1, 2, \cdots, n,$$

$$\zeta_j \geq \beta_j, \zeta_j \geq -\beta_j;\ j = 1, 2, \cdots, p.$$

- **Iterative coordinate descent algorithm**:
  Peng, B. and Wang, L. (2014) An iterative coordinate-descent algorithm for high-dimensional nonconvex penalized quantile regression. To appear in Journal of Computational and Graphical Statistics.

A numerical example

We first generate $(\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)^T$ from the multivariate normal distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\sigma_{jk})_{p\times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. The next step is to set $X_1 = \Phi(\tilde{X}_1)$ and $X_j = \tilde{X}_j$ for $j = 2, 3, \cdots, p$. The scalar response is generated according to the heteroscedastic location-scale model:

$$Y = X_6 + X_{12} + X_{15} + X_{20} + 0.7X_1\epsilon,$$

where $\epsilon \sim N(0, 1)$ is independent of the covariates.

A numerical example (cont'd)

We consider the following criteria.

Size: the average number of non-zero regression coefficients $\widehat{\beta}_j \neq 0$ for $j = 1, 2, \cdots, p$;

P1: the proportion of simulation runs including all true important predictors, namely $\widehat{\beta}_j \neq 0$ for any $j \geq 1$ satisfying $\beta_j \neq 0$. For the LS-based procedures and conditional median regression, this means the percentage of times including X_5, X_{12}, X_{15} and X_{20}; for conditional quantile regression at $\tau = 0.3$ and $\tau = 0.7$, X_1 should also be included.

P2: the proportion of simulation runs X_1 is selected.

AE: the absolute estimation error defined by $\sum_{j=0}^{p} |\widehat{\beta}_j - \beta_j|$.

A numerical example (cont'd)

Table: Simulation results ($n = 300$, $p = 600$)

Method	Size	P1	P2	AE
LS-Lasso	24.30 (0.61)	100%	7%	1.40 (0.03)
Q-Lasso ($\tau = 0.5$)	25.76 (0.94)	100%	10%	1.05 (0.03)
Q-Lasso ($\tau = 0.7$)	32.74 (1.22)	90%	90%	1.78 (0.05)
LS-ALASSO	4.68 (0.08)	100%	0%	0.37(0.02)
Q-Alasso ($\tau = 0.5$)	4.53 (0.09)	100%	0%	0.18 (0.01)
Q-Alasso ($\tau = 0.7$)	6.19 (0.16)	100%	86%	0.62 (0.01)
LS-SCAD	6.04 (0.25)	100%	0%	0.38 (0.02)
Q-SCAD ($\tau = 0.5$)	6.14 (0.36)	100%	7%	0.19 (0.01)
Q-SCAD ($\tau = 0.7$)	9.97 (0.54)	100%	100%	0.38 (0.03)
LS-MCP	5.56 (0.19)	100%	0%	0.38 (0.02)
Q-MCP ($\tau = 0.5$)	5.33 (0.23)	100%	3%	0.18 (0.01)
Q-MCP ($\tau = 0.7$)	7.56 (0.32)	98%	98%	0.37 (0.03)

Conclusion remarks

- **Quantile regression is useful for analyzing skewed, heteroscedastic expenditure data**
- **Quantile regression is useful for modeling high-dimensional heterogeneous data**

Resources and references

- Koenker, R. (2005) Quantile regression. Cambridge University Press.
- He, X., Wang, L. and Hong, H. (2013) Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. To appear in Annals of Statistics.
- Maidman, A. and Lan Wang. (2015) Predicting high spending customers using semiparametric quantile regression, technical report.
- Peng, B. and Wang, L. (2014) An iterative coordinate-descent algorithm for high-dimensional nonconvex penalized quantile regression. To appear in Journal of Computational and Graphical Statistics.
- Sherwood, B., Wang, L. and Zhou, A. (2014) Weighted quantile regression for analyzing health care cost data with missing covariates. Statistics in Medicine.
- Wang, L., Wu, Y. C. and Li, R. (2012) Quantile regression of analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*.

Questions/Comments

Contact information: Dr. Lan Wang
School of Statistics
University of Minnesota
wangx346@umn.edu